

# NLP Methods are Sensitive to Sub-Clinical Linguistic Differences in Schizophrenia Spectrum Disorders

Sunny X. Tang, M.D. – 汤小菁

Assistant Professor of Psychiatry – 精神病学学科 副教授

LDC-China Workshop

November 9-10, 2020



# Collaboration



Mark Liberman  
Sunghye Cho  
Reno Kriz



Dan Wolf  
Raquel Gur  
Suh Jung Park



João Sedoc



Mahendra Bhati

- This project was a collaborative effort across institutions



# Conflicts of Interest - Disclosures



Research support and consulting



**North Shore**  
— THERAPEUTICS —

Co-Founder and Equity



**Neurocrine**<sup>®</sup>  
B I O S C I E N C E S

Consultant

- I have some industry-related financial conflicts of interest



# Outline

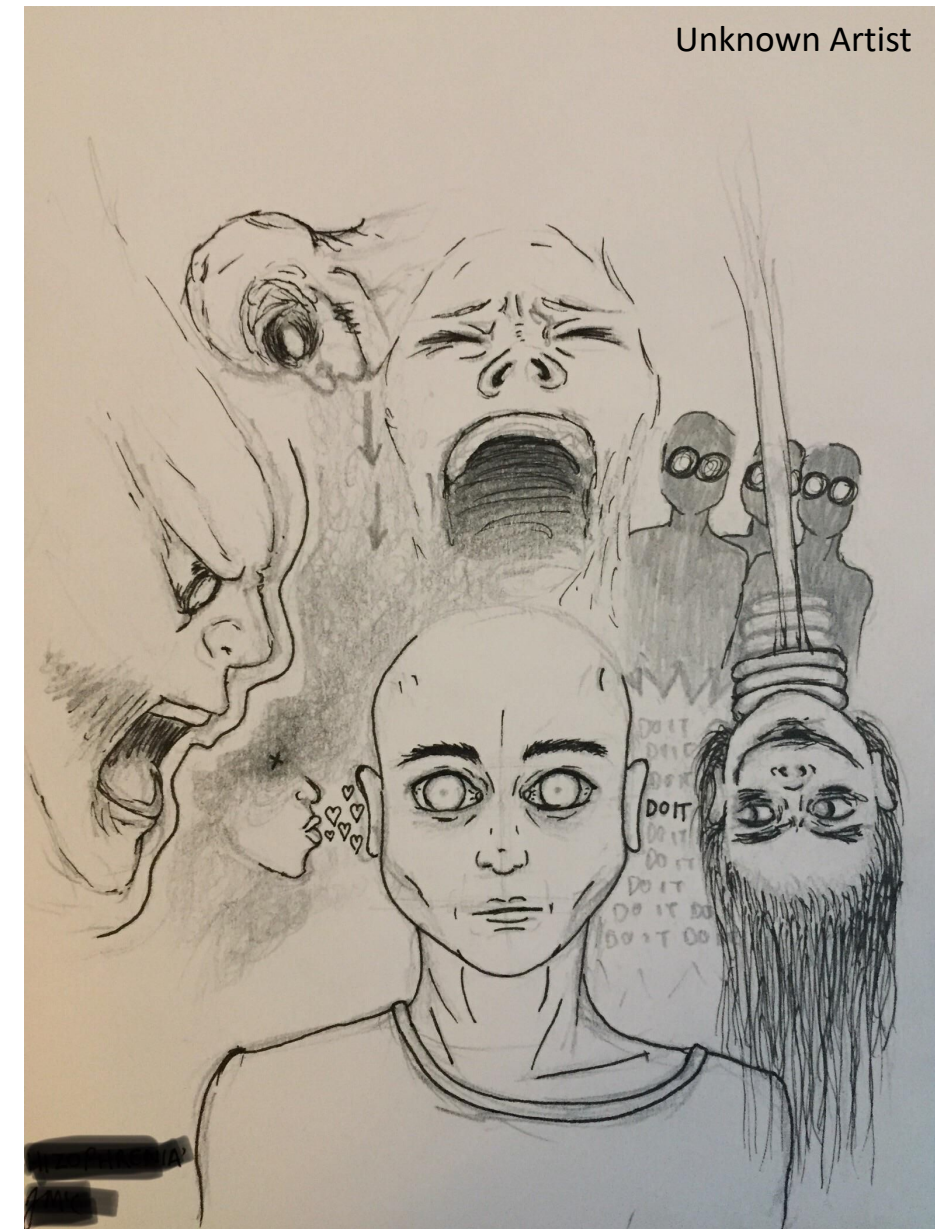
- ▶ What is schizophrenia and why should we care about language?
- ▶ Pilot study: Exploration of NLP methods
  - Words
  - Parts-of-speech
  - Sentence-level
  - Prediction Models
- ▶ New directions





## Core Symptoms of Schizophrenia

- ▶ Hallucinations – auditory, visual or other sensory experiences like hearing voices
- ▶ Delusions – fixed false beliefs like paranoid thoughts of people being out to hurt you
- ▶ Disorganized behaviors – actions that don't make sense, like wearing heavy clothing in the summer
- ▶ Avolition – decreased motivation
- ▶ Asociality – decreased interactions with others
- ▶ Anhedonia – decreased enjoyment
- ▶ Cognitive impairment – difficulty with attention, memory, social cognition and other brain functions



- Schizophrenia is defined by a constellation of symptoms, but not every symptom is present in every patient

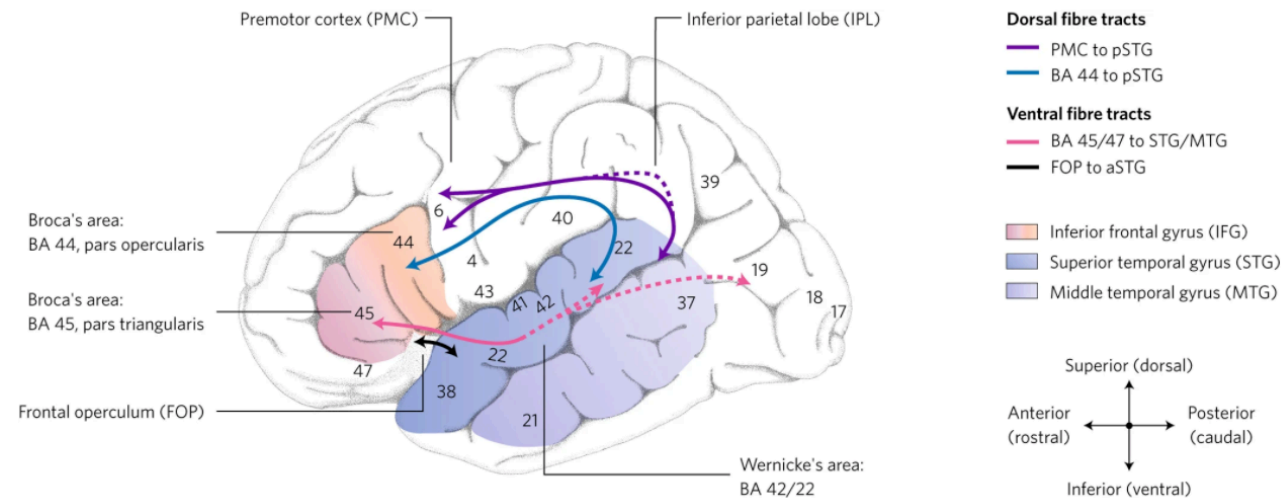


# Language Disturbance in Schizophrenia

- ▶ **Incoherence and Disorganization** – while appropriate words are used with normal grammar, spoken and written language “don’t make sense”
- ▶ **Poverty of Speech and Content** – total speech amount is decreased or very little meaning is conveyed
- ▶ **Unique Features** – e.g. neologisms, echolalia, clanging
- ▶ **Excitement** – fast, pressured, or increased speech quantity

**Fig. 1: Structural connectivity between language regions.**

From: *Language, mind and brain*



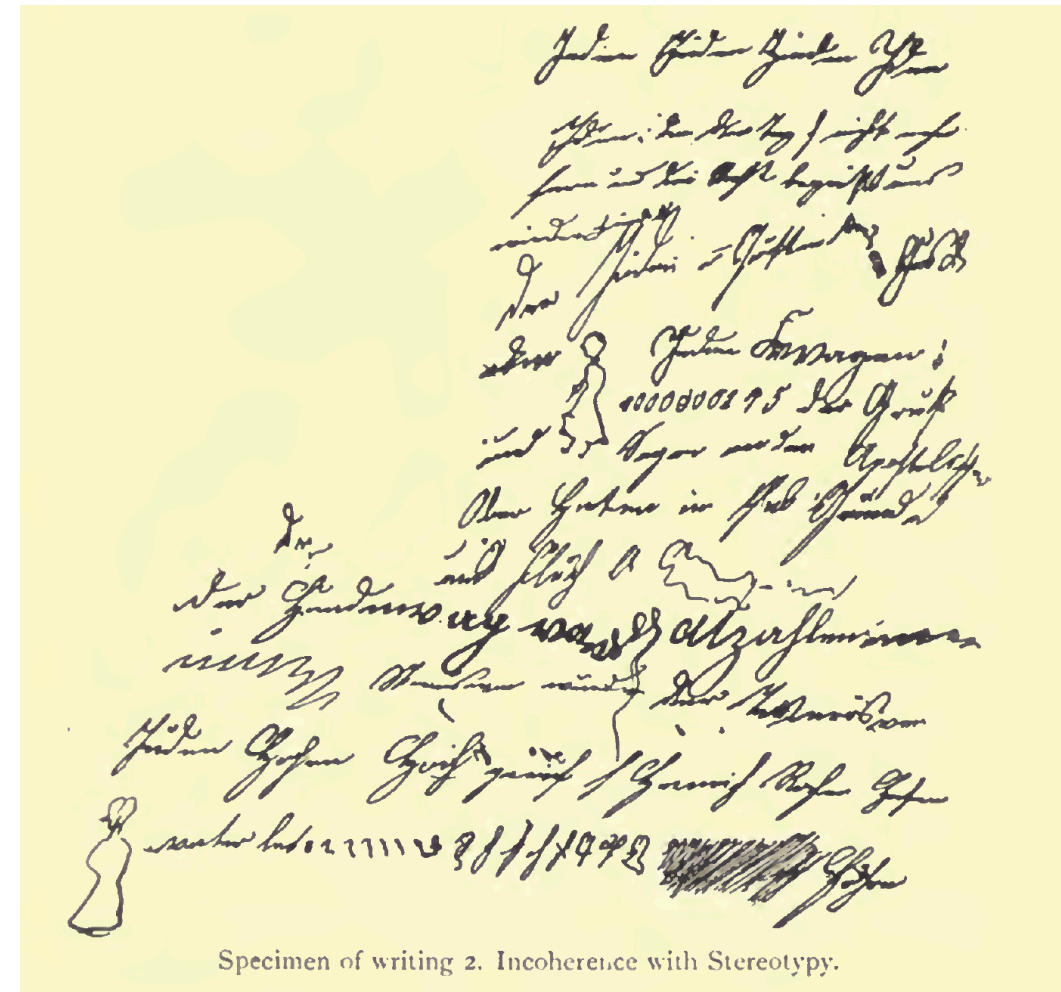
Friederici et al, 2017 *Nature Human Behavior*

- Language disturbance is a hallmark of schizophrenia – likely reflecting changes in brain circuitry



# Language Disturbance in Schizophrenia

- ▶ **Examples:**
- ▶ Then I left San Francisco to. Where did you get that tie?
- ▶ I like the war weather in San Diego. Is that a conch shell on your desk?
- ▶ It happened in eons and eons and stuff they wouldn't believe in him. The time that Jesus Christ people believe in their thing people believed in, Jehovah God that they didn't believe in Jesus Christ that much.
- ▶ Parents are the people that raise you. Anything that raises you can be a parent. Parents can be anything, material, vegetable, or mineral, that has taught you something.



Kraepelin, *Dementia Praecox*





# Schizophrenia Basics

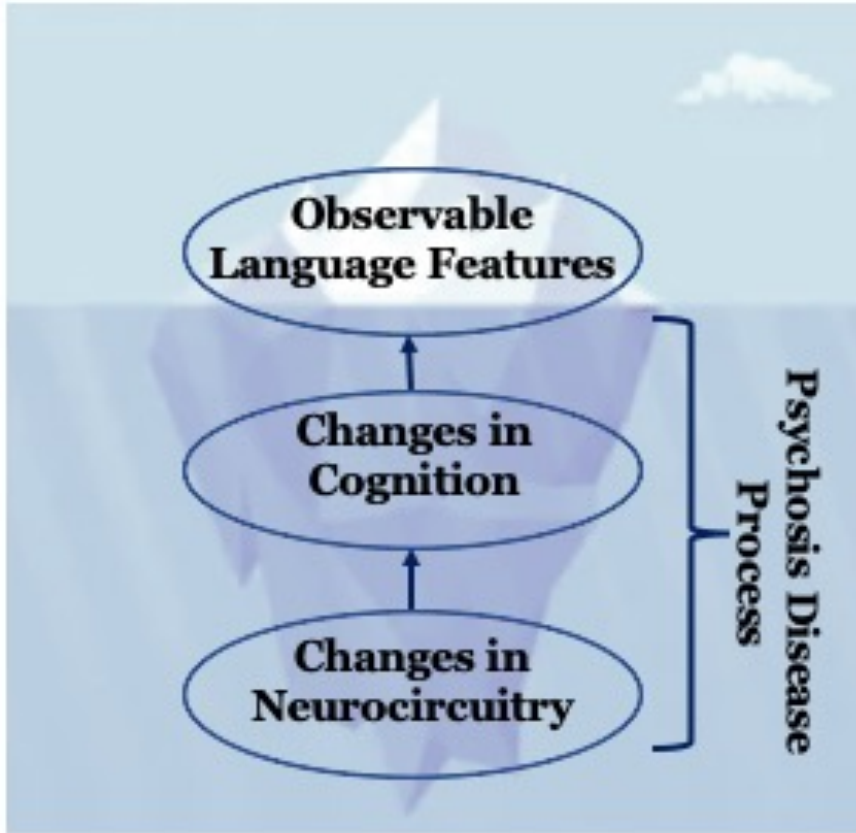
- ▶ Population prevalence approximately 1%
- ▶ One of the leading causes of disability worldwide – onset in adolescence
- ▶ Economic burden is \$115 billion per year in the US alone
- ▶ Life expectancy is 20 years less than for general population
- ▶ Yet! Possibility of prevention and successful treatment
- ▶ Dire need for better biomarkers for:
  - Prediction
  - Diagnosis
  - Tracking treatment response
  - Tailoring treatment

- Schizophrenia is a devastating illness that affects 20 million people worldwide





# Rationale for Studying Language in Schizophrenia



**Fig 1:** Model - Language as observable reflection of underlying psychosis disease processes


- ▶ Easily observable
- ▶ Direct reflection of brain processes
- ▶ NLP:
  - Objective
  - Automated
  - Scalable

Language is a valuable biomarker that can transform the way we diagnose, treat, and do research in schizophrenia

- Speech is the observable surface phenomenon that reveal the inner mind



# Outline

- ▶ What is schizophrenia and why should we care about language?
- ▶ Pilot study: Exploration of NLP methods 
  - Words
  - Parts-of-speech
  - Sentence-level
  - Prediction Models
- ▶ New directions



# Pilot Study – Sample

	Healthy Control Participants	Schizophrenia Spectrum Disorder Participants	p value	Cohen's d
	HC	SSD		
<b>Sample n</b>	11	20		
<b>Cohort</b>			0.10	
Cohort 1	5	15		
Cohort 2	6	5		
<b>Age (mean years ± SD)</b>	35.6 ± 5.8	36.5 ± 7.2	0.75	0.12
<b>Sex (n, %)</b>				
Female	7 (64%)	9 (45%)	0.32	
Male	4 (36%)	11 (55%)		
<b>Race (n, %)</b>			0.12	
African American	3 (30%)	13 (65%)		
Asian	0 (0%)	1 (5%)		
Caucasian	7 (70%)	6 (30%)		

**\*NOT enriched for "thought disorder"!**

- Pilot sample collected from 2 cohorts at Penn – not enriched for speech disturbance



# Dimensions of Speech Disturbance in Schizophrenia

## ▶ Disorganization

- Derailment
- Tangentiality
- Incoherence
- Illogicality
- Circumstantiality
- Loss of Goal

## ▶ Negative/Poverty Features

- Poverty of Speech
- Poverty of Content of Speech
- Perseveration
- Increased Latency
- Decreased Intonations / Flattening

## ▶ Idiosyncratic / Semantic

- Stilted speech
- Word approximations
- Neologisms
- Clanging

VOL. 12, NO. 3, 1986

### **Scale for the Assessment of Thought, Language, and Communication (TLC)**

by *Nancy C. Andreasen*

The following set of definitions was  
developed to improve the reliability

or perceptual disorders) as manifesta-  
tions of their schizophrenia.

- Clinical rating scale for speech disturbance in schizophrenia identifies 18 items





# Pilot Study – Clinical Ratings Details

	HC	SSD	p value	Cohen's d
Items: Mean (SD)				
1. Poverty of Speech	0.00 (0.00)	0.10 (0.31)	0.29	0.40
2. Poverty of Content of Speech	0.00 (0.00)	0.25 (0.44)	0.07	0.70
3. Pressure of Speech	0.00 (0.00)	0.10 (0.45)	0.47	0.28
4. Distractible Speech	0.00 (0.00)	0.00 (0.00)	1.00	0.00
5. Tangentiality	0.27 (0.65)	0.20 (0.89)	0.81	0.09
6. Derailment	0.00 (0.00)	0.20 (0.62)	0.29	0.40
7. Incoherence	0.00 (0.00)	0.25 (0.64)	0.21	0.48
8. Illogicality	0.00 (0.00)	0.30 (0.73)	0.19	0.51
9. Clanging	0.00 (0.00)	0.05 (0.22)	0.47	0.28
10. Neologisms	0.00 (0.00)	0.10 (0.31)	0.29	0.40
11. Word Approximations	0.00 (0.00)	0.20 (0.52)	0.22	0.47
12. Circumstantiality	0.18 (0.40)	0.25 (0.72)	0.77	0.11
13. Loss of Goal	0.00 (0.00)	0.10 (0.45)	0.47	0.28
14. Perseveration	0.00 (0.00)	0.05 (0.22)	0.47	0.28
15. Echolalia	0.00 (0.00)	0.00 (0.00)	1.00	0.00
16. Blocking	0.00 (0.00)	0.10 (0.31)	0.29	0.40
17. Stilted Speech	0.00 (0.00)	0.10 (0.45)	0.47	0.28
18. Self-Reference	0.00 (0.00)	0.15 (0.49)	0.32	0.38

- Largest effect sizes in poverty of content > Illogicality ~ Incoherence ~ Word Approximations



# Pilot Study – NLP Methods

## ➤ Individual Words



- ▶ Verbatim transcription of recordings, *including disfluencies*
- ▶ Odds Ratio calculated
- ▶ Log transformed
- ▶ Weighted based on informative Dirichlet prior (relative to expected frequency)

- Word-level analysis compared weighted log-transformed odds for individual words





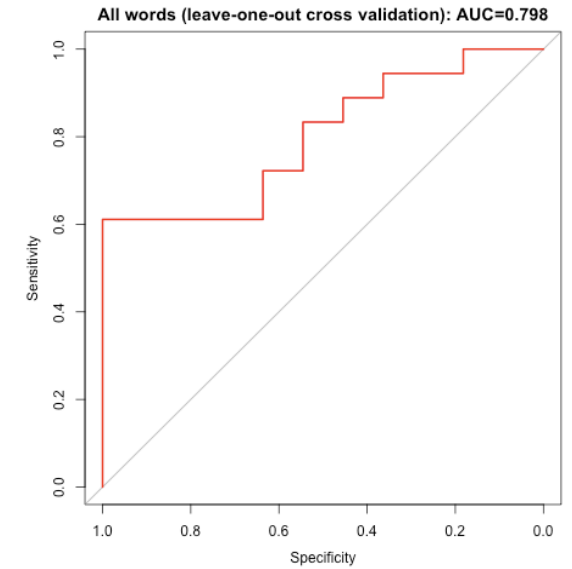
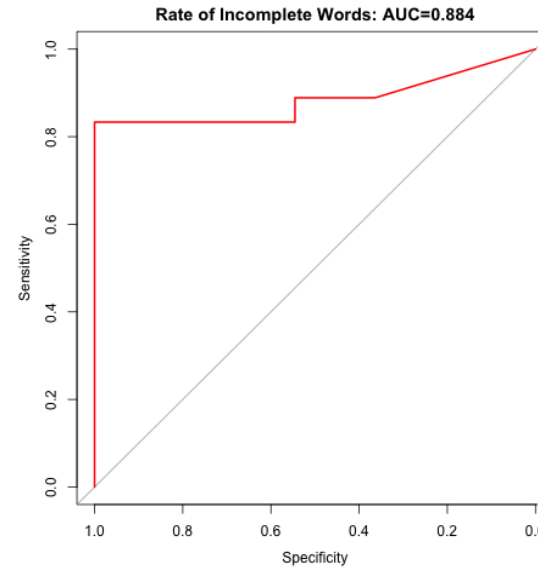
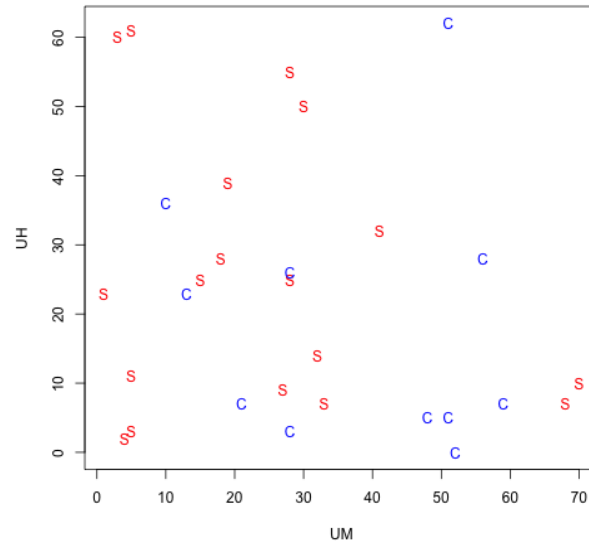
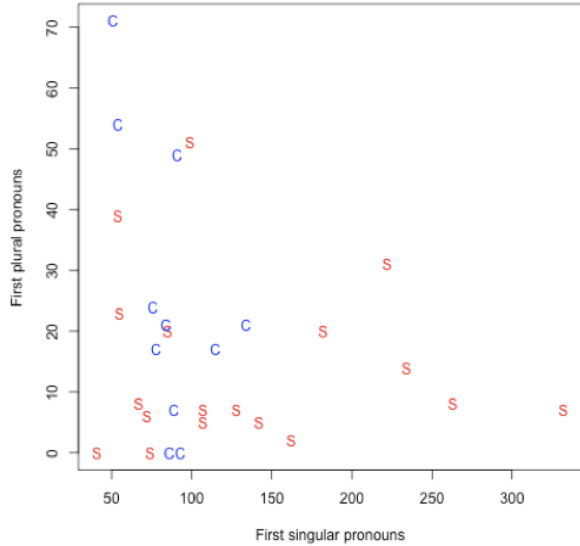
# Pilot Study – Word Frequency

Top SSD-Associated Words				Top HC-Associated Words			
Token	SSD Frequency	HC Frequency	Weighted Log-Odds	Token	SSD Frequency	HC Frequency	Weighted Log-Odds
[FPS] I/me...	94.7	60.7	7.2	um	16.3	25.6	-3.8
he	7.3	1.3	4.7	[FPP] we/us	9.9	17.6	-3.8
[Incomplete Word]	4.6	1.0	3.5	like	16.1	24.6	-3.4
they	8.2	4.3	2.7	of	10.6	17.0	-3.2
no	3.6	1.2	2.5	actually	0.6	2.4	-2.8
[Identifying Name]	3.8	1.5	2.3	[Laughter]	0.6	2.3	-2.7
[SP] you/your...	18.7	13.7	2.3	so	10.0	15.3	-2.7
lived	1.4	0.2	2.1	sort	0.1	1.2	-2.6
uh	17.3	12.4	2.1	usually	0.3	1.7	-2.6
well	3.9	1.7	2.1	ago	0.3	1.7	-2.5
used	2.1	0.6	2.1	great	0.3	1.3	-2.2
on	6.5	3.9	2.0	awesome	0.0	0.6	-2.2
cause	1.7	0.5	1.9	super	0.0	0.6	-2.2
him	1.4	0.4	1.9	bunch	0.0	0.6	-2.0
know	13.2	9.9	1.8	as	1.6	3.2	-2.0
people	2.7	1.2	1.8	gone	0.0	0.5	-2.0
never	1.7	0.7	1.7	wife	0.0	0.5	-2.0
had	4.5	2.6	1.6	places	0.1	0.8	-2.0
mom	1.7	0.7	1.6	recently	0.1	0.7	-2.0
florida	0.6	0.1	1.6	definitely	0.0	0.6	-1.9

- Distinct patterns in word usage in SSD vs. HC



# Pilot Study – Word Frequency



FPS pronoun use in SSD  
vs.  
FPP pronoun use in HC

“Uh” in SSD  
vs.  
“Um” in HC

Incomplete words  
predict SSD group  
AUC = 0.88

Overall word usage  
predict SSD group  
AUC = 0.80

- First person pronoun use, filler word use, incomplete words distinguish SSD from HC speech



# Pilot Study – NLP Methods

## ➤ Parts of Speech

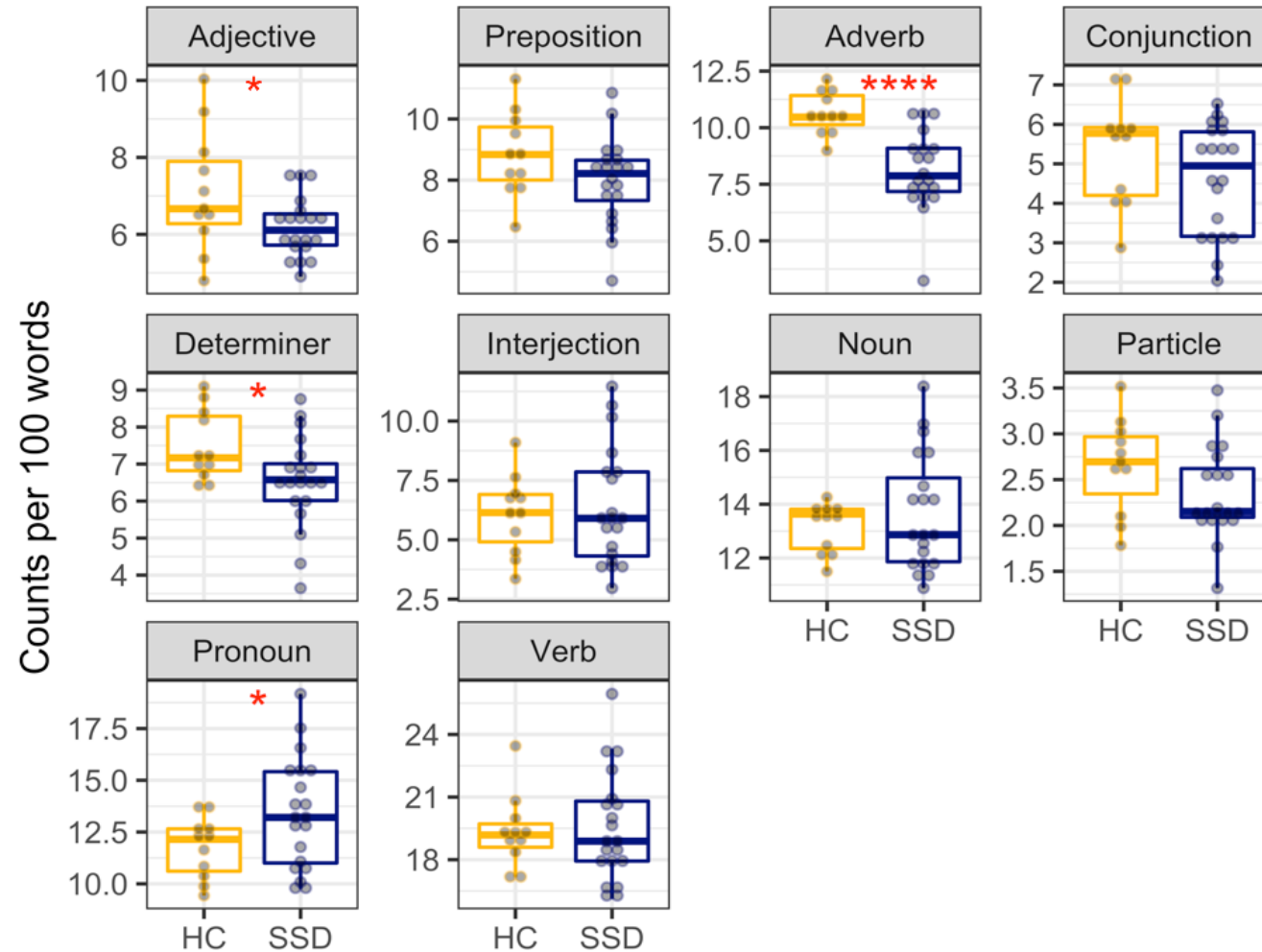
enormous <sup>DISEASE</sup> quantities of diverse, disconnected data. These data sets present substantial analytic challenges, but can also illuminate new <sup>DRUG</sup> avenues of <sup>DRUG</sup> inquiry that yield unprecedented improvements in global health. Roam is realizing this <sup>DISEASE</sup> potential by combining our <sup>DISEASE</sup> proprietary data platform with <sup>DRUG</sup> advanced machine <sup>DISEASE</sup> learning, empowering life sciences companies, <sup>DISEASE</sup> hospital systems, insurers, and

- ▶ Tokenized using spaCy using basic model 'en\_core\_web\_sm'
- ▶ Compared token counts covarying for age, sex, cohort, education level

- POS analyses compared spaCy-defined parts of speech characteristics



# Pilot Study – Parts of Speech

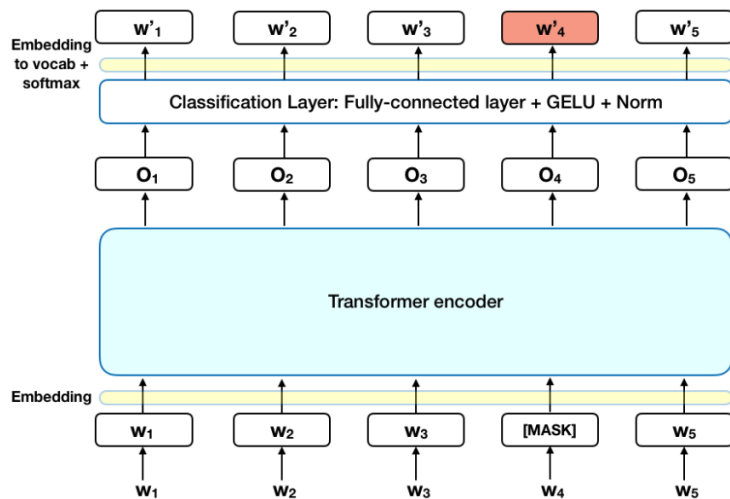


- People with SSD used fewer adjectives, adverbs, and determiners – more pronouns



# Pilot Study – NLP Methods

## ➤ Sentences



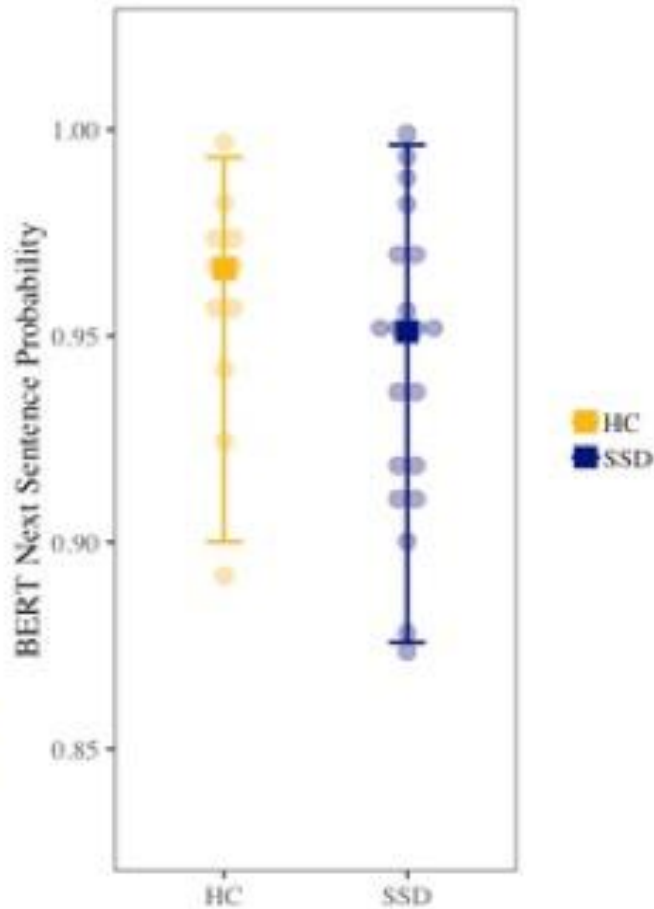
- ▶ Tokenized using NLTK using intuitive punctuation points defined by transcribers
- ▶ BERT next-sentence predictability
- ▶ BERT embedding distances by turn
- ▶ NOTE! Initial failure with using spaCy for sentence-tokenization – picked up disfluency
  - E.g. “I sto- // stopped at the store”
  - E.g. “She gave // gave me the flowers”

- Sentence-level approach leveraged BERT model



# Pilot Study – Sentence Level Analysis with BERT

C) BERT Next-Sentence Probability



## Examples from TLC:

Then I left San Francisco to. Where did you get that tie?

BERT = 0.0053

I like the war weather in San Diego. Is that a conch shell on your desk?

BERT = 0.0179

It happened in eons and eons and stuff they wouldn't believe in him.

BERT = 0.9999

The time that Jesus Christ people believe in their thing people believed in, Jehovah God that they didn't believe in Jesus Christ that much.

Parents are the people that raise you.

BERT = 0.9999

Anything that raises you can be a parent.

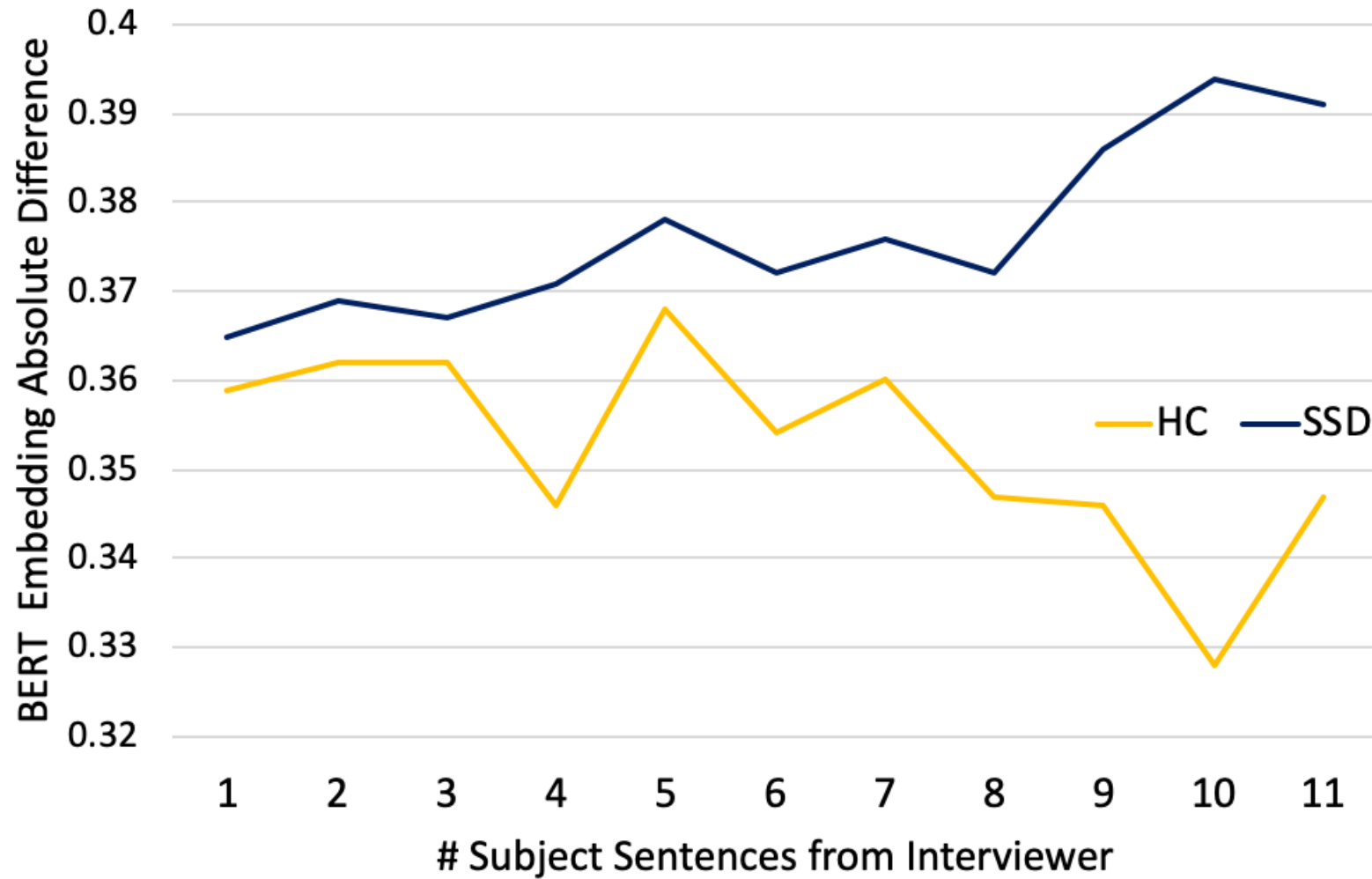
Parents can be anything, material, vegetable, or mineral, that has taught you something.

BERT = 0.9999

- Next sentence probability – some separation between groups, but not significant.



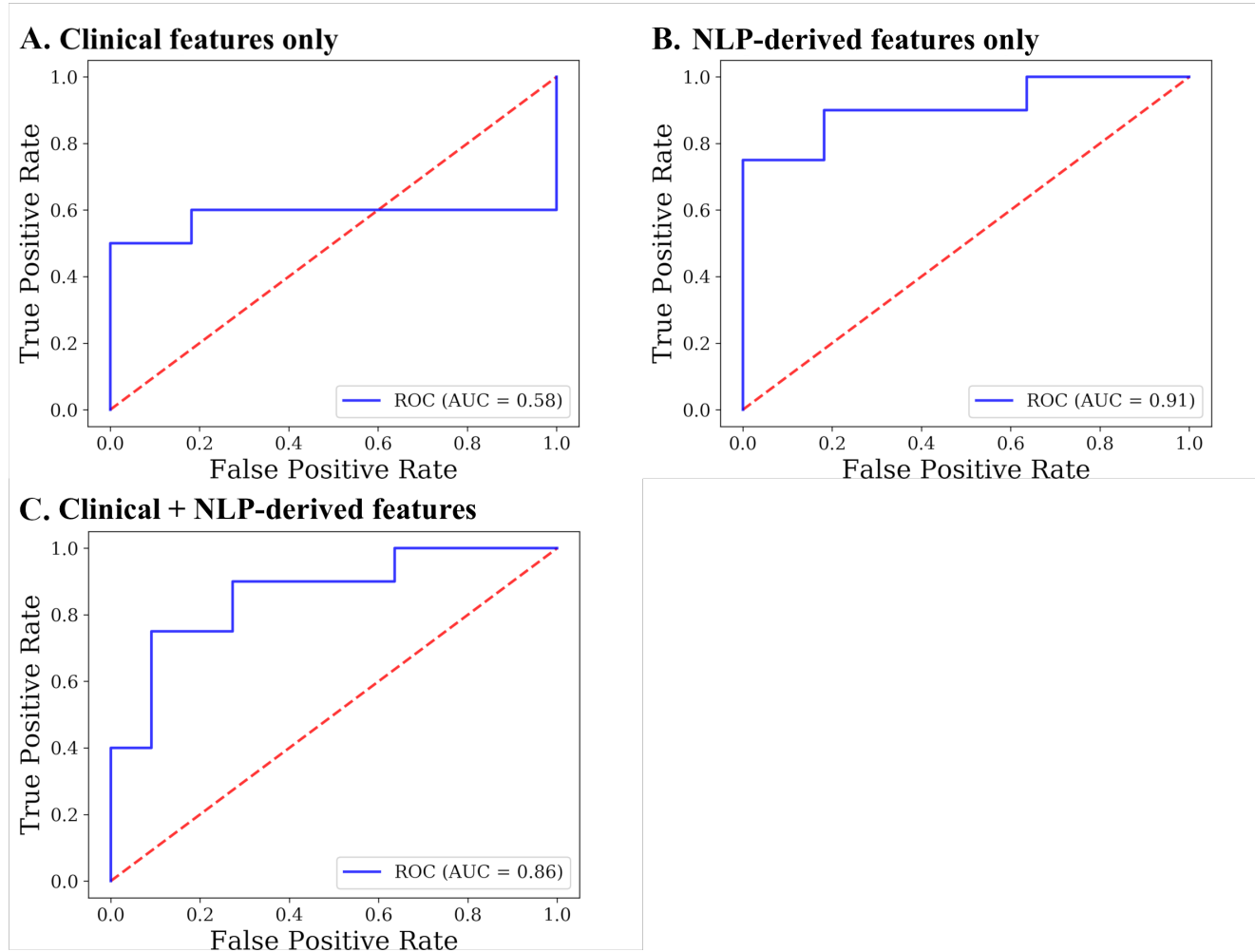
# Pilot Study – Sentence Level Analysis with BERT








# Pilot Study – Predicting SSD Diagnosis





# Outline

- ▶ What is schizophrenia and why should we care about language?
- ▶ Pilot study: Exploration of NLP methods
  - Words
  - Parts-of-speech
  - Sentence-level
  - Prediction Models
- ▶ New directions 



# Future Directions

## Questions

- ▶ What about character-based languages like Chinese?\*
- ▶ Does this replicate?
- ▶ Which NLP measures correspond to which clinical features?
- ▶ What about predicting specific psychosis symptoms?
- ▶ Are there language changes that tell us if a person is likely to develop psychosis in the future?
- ▶ Are there language changes that tell us whether a patient will respond to medication and which medication?
- ▶ What do language differences imply about how people with SSD think about others and themselves?
- ▶ What about other disorders?

## Next Steps

- ▶ Cross-cultural dataset
- ▶ Large(r) prospective study
- ▶ Collect information on symptom dimensions
- ▶ Collect information on social cognition
- ▶ Large screening of many people with different disorders
- ▶ Investigate neuroimaging changes and their relationship to language

\*王久菊, 王鹏飞, 权文香, 田菊, 刘津, 董问天, 精神分裂症的语言认知特点及其脑机制, 生物化学与生物物理进展, 2015, 42(1) : 49-55.



Questions?

*thankyou*

謝謝

# Pilot Study – Parts of Speech Characteristics

